

Data Analysis

Lon Cardon
Ellen Wijsman
Marcella Devoto
<Teri Manolio>

Data analysis status

- Combination of
 - (1) genome sequence+hapmap;
 - (2) large samples;
 - (3) genotyping throughput/cost;
 - (4) rigorous inference

have led to unprecedented numbers of new genes/regions.

- Findings are
 - Unequivocal and reproducible.
 - Mostly moderate to small effects.
- There are more discoveries than we can handle.
- There are many more loci yet undiscovered.

Data analysis challenges

1) software

- Legacy software in statgen is serious hindrance to engaging expertise of others with potentially valuable contributions
 - Getting a bit better (PLINK, R), but still focused inward
- Cf: bioinformatics as discipline has 1/10 history of quantitative genetics in time, yet is already ahead in software

Data analysis challenges:

2) statistical methods

- Stat gen success cannot be attributed to analysis methods
 - Design, interpretation have contributed a lot, but analysis methods behind most findings very simple.
 - Early days for GxG, GxE methods – much more to do with the data already collected
 - Do all genetics applications fit into inferential statistics framework? Cf microarray field

Data analysis challenges

3) causality & functional analysis

- Identifying causal variants
 - GWAS do not identify causal variants.
 - Statistics and strong study design can help (see recent T1D/MHC Nature), but ultimate proof comes from other studies
- What to do with the information we have just obtained?
 - Solution (partial): study them as deeply as possible genetically before initiating long-term functional experiments.
 - There are methodology implications with this (e.g, how to deal with resequencing data...)

Heterogeneity

- Accommodating, understanding, embracing heterogeneity
 - Genetic, allelic, population/ethnic, phenotypic
- ***Heterogeneity is a bonus, not a nuisance.***
 - This is a strength of US.
 - Cross-ethnic fine-mapping, selection methods, admixture, ...
 - Need to support research on combined data, methods, ideas to exploit this attribute.
- A lot of resources going into assoc right now. Need to maximize potential by combining studies.
 - ‘Combining’ starts with making data available to all, but requires much more
 - Need collaboration at level of phenotype, analysis, design (replication), molecular follow-up, ..., translation

Data analysis – making use of all skills

- At present, almost all analyses done in academic setting
- Others with important skills to contribute:
 - Pharma Industry
 - Computing/software Industry
 - FDA/EMEA
 - NIH/NHGRI staff
 - Need to find ways to increase the number of people at the table.

Genetic analysis: where are we going from here?

1. More of same with new diseases/traits
 - ...to find new loci for new traits
2. Find more variants in diseases already studied
 - Larger samples, meta-analyses
3. Figure out what to do with the loci already found

Needs

- Applications/analyses
 - Studies without new data (combining data)
- Methods
 - GxE, GxG, CNVs, rare variants
 - Integration with –omics data
- Software
- Training
- Bringing together people from different disciplines